Computer Vision and Image Understanding 156 (2017) 1–3

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Editorial

Image and Video Understanding in Big Data☆

The huge volume of data produced every day is posing a significant challenge to computer scientists since it is infeasible that this data can be effectively processed and consistently interpreted by humans manually, even to a very small extent. Due to the automation of many industrial processes and the advent of cheaper and high-performance sensors, many aspects of life, including medical, commercial, industrial and security areas are increasingly more characterized by large data collections that must also be processed and communicated with. DNA sequencing, radiographic and imaging examinations (e.g. MRI, X Ray, echography), urban surveillance, traffic monitoring, customer profiling and e-commerce, visual inspection, industrial machine maintenance and failure prediction are some examples of the vast range of possible applications in our everyday social and working life characterized by large data gatherings.

Among these, visual data takes a prominent role given the large and pervasive diffusion of imaging devices in our cities, industries, at home, and in our hands given the plenitude of personal devices such as smart phones at our disposal. For the latter, one can appreciate readily the scale of the data consumed just by thinking about the amount of image and video data downloaded every minute in social media such as Facebook, Instagram, Snapchat, among others. Moreover, visual data is by large the most diverse and demanding media, as compared to text for instance, growing at an unprecedented speed. This requires the design of effective methods to manage it, by mining relevant information while discarding redundant or useless data. To that end, more scalable and robust methods are required to efficiently index, retrieve, organize, interpret and interact with such big visual data, and this cannot be done without automatic or semi-automatic, e.g. human in the loop, methods and processes capable of distil useful observations from a large quantity of raw data.

In this context, it is clear that big visual data analysis and understanding impose significant scientific and technological challenges since one not only should have to advance methods able to gracefully scale to both big and diverse data whilst being computationally cheap, but also should need to develop processes suitable for learning from big and diverse data *without* incurring prohibitive costs in human and monetary resources, and of time, required by exhaustive data annotation. Moreover, efficient *online learning* methods could be required to cope with data acquired over time.

This special issue brings together fourteen selected papers that reflect some of the current trends, progress and challenges from big visual data analysis and understanding.

Most of the big visual data available refers to *human* data. Whole human body or part of it (e.g., upper body) and faces especially, constitute a huge amount of available data from which interesting information can be extracted.

Castrillón-Santana et al. focus on the problem of gender classification from faces in the wild by adopting a multi-expert approach. A set of experts are trained using several combinations of local operators to capture and fuse different aspects of the face images. Feature-level and score-level fusion are explored for the final classification. Besides, it is also demonstrated that the score-level fusion approach followed by a further classification stage on the score results showed reasonable performance as a reasonable computational cost, proposing it as a good trade-off between accuracy and cost.

Lo Presti and La Cascia address face emotion recognition via analysing dynamics of face expression. Given a temporal sequence of face images, several descriptors are extracted per frame, and a set of Hankel matrices per feature-scale pair are estimated on top for modelling the face dynamics. Such matrices, after a random subspace projection, are utilized within a boosting approach to build a strong classifier, reaching state-of-the-art performance on benchmark datasets. The same approach is shown to work satisfactorily for pain detection.

Segalin et al. explore a novel problem of large-scale image understanding with the goal to discover persons' personality traits based on the images the person preferred and liked on social networks. They demonstrate the viability of using deep convolutional neural networks for inferring five personality traits, including openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. This type of research opens new possibilities of profiling users and suggesting friends with common personality traits on social networks.

Human identification is undoubtedly a central task in computer vision and a fundamental problem to be addressed in many applications, video surveillance *in primis*. As such, the huge amount of data acquired by surveillance cameras constitutes a big data analysis challenge that needs to be met effectively. Paisitkriangkrai et al. study the effectiveness of learning person re-identification (re-id) matching similarity distance metric ensembles from multiple low-level visual features. They examine the idea of optimising directly person re-id testing evaluation values given by the Cumulative Matching Characteristic (CMC) curve on benchmark datasets.

☆ The guest editors would like to thank Dr. John Smith for his initial contribution to the organization of this special issue.

http://dx.doi.org/10.1016/j.cviu.2017.01.007
1077-3142/

They show that maximising the correct identification among the top ranked candidates in CMC is more beneficial than maximising the relative distance between a matched pair and a mismatched pair, suggesting optimising a matching objective against the actual testing pool criteria rather than discriminating the probe itself is more effective in yielding a better matching distance metric for person re-identification.

Das et al. address the image-based person identification/recognition problem casted as the assignment of an identity to the image of an individual based on a model learned on his or her appearance. The method consists in a convex optimization-based iterative strategy able to select, for annotation, a sparse set of non-redundant representative training images progressively in an online setting from multi-sensor data. In this way, the human labeling effort is reduced while maintaining good identification accuracy, which is also increased whenever more data becomes available. This framework finally uses a structure preserving sparse reconstruction based classifier to reduce the training burden while enabling an online update of the identification framework involving only new samples, without the need to train from scratch whenever new batch of data arrives. Overall, the proposed framework uses two convex optimization based strategies to select a few informative but non-redundant samples for labeling and to update a person identification model online. Experiments on three publicly available benchmark datasets for re-identification are performed to validate the proposed approach.

Recognition of activities is another issue extensively addressed by the computer vision community due to its critical role in many applications, e.g. smart city management and video surveillance being two of the primary areas. Richard and Gall tackle the problem of learning bag-of-feature representation for action recognition, which suffers from a lack of discriminative power, since it is not usually optimized jointly with the classifier. The authors propose a recurrent neural network that is equivalent to the traditional bag-of-words approach which enables discriminative training. The model further allows to incorporate the kernel computation into the neural network directly, solving the complexity issue of computing kernels.

The work of Vaca-Castano et al. addresses a specific activity recognition task, dealing with egocentric vision of daily activities. This type of data is characterized by a few prototypical scenes regardless of the actors who are performing the activity, and the work just aims at proving as scene identification can be improved by using temporal context. Besides, it also copes with object detection showing how generic object detection can be improved when taking into consideration the scene label, in other words, by rescoring the object detection results according to the scene content. For this task, two algorithms are proposed: in the supervised case, when the labels of the test videos are explicitly predicted from scene models learned in training data, a greedy algorithm and a Support Vector Regression based method were devised. In the unsupervised case, a formulation based on Long Short-Term Memory (LSTM) directly infers the probability of having a type of object in a sequence, without an explicit knowledge of the label of the scenes.

For big visual data analysis, a model is often required to cope with large scale data points search and matching in very fast time. Hashing is one way to address this problem. A majority of recent studies on hashing addresses it in static image collections. Given the extra temporal information, videos are typically represented by features in much higher dimensional spaces compared to those representing images. This high dimensionality causes computational complexity problems for conventional hashing methods. To that end, Qin et al. studies a hashing method that aims to learn similarity-preserving binary codes by exploiting the correlations among different feature dimensions. This is designed for efficient nearest neighbour search in high-dimensional video data. The effectiveness of the model is demonstrated on action retrieval from video databases.

Anomalous event detection in video is another important task in intelligent video surveillance. Conventional approaches to automatic analysis of complex video scenes typically rely on handcrafted appearance and motion features, without the consideration of the visual context from which activities arise. To learn event descriptors specific to the visual context of interest, Xu et al. design an Appearance and Motion DeepNet model to exploit automatically the complementary information of appearance and motion in context, with a two-staged fusion strategy. Specifically, stacked denoising auto-encoders are deployed to learn appearance and motion feature representations. Multiple one-class SVM models are trained from these deep features to predict the anomaly scores of each input. A decision fusion scheme is applied to combine the individual scores for abnormal event detection.

However, big training data is not always available for all the classes. In particular, in most cases not all the classes are evenly sampled in their training data. That is, there exists an imbalanced data problem, giving rise to a long tail distribution among the samples available across all classes. A possible way to mitigating this lack of training data for many classes of interest is by having a model to interact with the environment in order to collect sufficient volumes of images necessary for learning all the classes of interest. To that end, Malmir et al. use deep learning to address the problem of active object recognition, where an agent interacts with the environment in order to recognize the object of interest. In this setup, the recognition problem is combined with a data sampling selection problem, i.e. which action to apply next in order to enrich the most deprived classes. This approach aims to train a deep convolutional network in a Q-learning framework for joint prediction of the object label and the action.

Reliable object detection in large scale imagery data is challenging, especially in videos. To avoid the prohibitive costs from annotating training video data, weakly supervised learning for object detection has been gaining significant attention in recent years. Specifically, visually similar objects are extracted automatically from sparsely labelled videos, bypassing exhaustive manual labelling of the training data. However, this visual appearance similarity approach to weakly supervised learning does not fare well with small or medium sized objects when appearance or motion is unreliable. To address this problem, Srikantha and Gall exploit additionally weakly-labelled video information by observing human-object interaction, which characterises an object's functionality, more than its appearance and motion. They demonstrate that object models trained using this approach to weakly supervised learning can yield between 86% and 92% of the performance given by their fully supervised counterparts.

Recognition and classification are essential for processing big data. In this context, Mai et al. present a label tree based method for efficient large-scale multi-class image classification, which is particularly suitable when the number of classes is very large. The proposed method learns effective and balanced trees by jointly optimizing balance and confusion (similar classes) constraints, so as to reduce the complexity. By organizing classes in a hierarchical structure, the number of classifier evaluations necessary for a test sample is reduced if a balanced tree is learnt. The proposed approach formulates the learning tree structure problem within an optimization framework in which the balance constraint is solved using integer linear programming and the confusion constraint is solved using a variant of k-means clustering. An extensive experimental phase is reported to compare the proposed method with other state-of-the-art techniques on four large-scale public datasets.

Cakir et al. propose an online supervised hashing technique for fast similarity image retrieval, which is indispensable when

dealing with large-scale data. Hashing is an historical, effective technique which can provide both fast search schemes and compact index structures, eventually mitigating the otherwise costly search procedure. Batch-learning strategies showed to be ineffective for large datasets, other than being not scalable in case of online, sequential data acquisitions. This work deals with such issues by presenting an online supervised hashing technique based on error correcting output codes. Assuming a stochastic setting where data arrives sequentially, this method learns and adapts its hashing functions in a discriminative manner, making no assumption about the number of possible class labels, and accommodating new classes as they are presented in the incoming data stream. In comparison with current approaches, this method shows a significant improvement in retrieval accuracy compared to state-of-the-art online hashing solutions, while being orders-of-magnitude faster than state-of-the-art batch methods.

It is evident that deep learning has been recently shown to be more attractive than many conventional machine learning methods to build recognition models from big data. Deep learning naturally benefits from big and diverse data sources, but it is also connoted by a significant caveat. Acquiring sufficiently labelled data for training deep neural networks can be both tedious and expensive, in terms of not only human resource and monetary costs, but also the cost of time. In this context, *transfer learning* (TL) methods can help to circumvent this problem, trying to exploit previous (source) knowledge to the (target) domain at hand. In this line, the work of Kuzborskij et al. address the object detection problem by a specific transfer learning approach, Hypothesis TL (HTL). This work considers pre-trained models as black boxes and face the HTL problem as the task of efficient selection and combination of source hypotheses from a large (of order of 10^3) pool. To this end, by leveraging on the subset selection problem from the literature, a greedy algorithm is introduced that attains the state of the art performance using a small amount of data from the target domain. Besides, a randomized approximated variant is also designed, which is independent from the number of sources, with no loss in performance.

In summary, this special issue presents a small collection of works that nevertheless highlight on how the research community is addressing the problem of big visual data analysis and understanding. Even though this is only a limited set of examples, the variety of the methods developed and the wide range of applications addressed undoubtedly demonstrate the growing importance of this field. This is surely to become increasingly one of the protagonists in the computer vision and machine learning communities in the coming years.

Guest Editors

Vittorio Murino
Shaogang Gong
Chen Change Loy
Loris Bazzani